Sparse Factorizations of Gene Expression Guided by Binding Data

L. Badea and D. Tilivea

Pacific Symposium on Biocomputing 10:447-458(2005)

SPARSE FACTORIZATIONS OF GENE EXPRESSION DATA GUIDED BY BINDING DATA

LIVIU BADEA and DOINA TILIVEA
AI Lab, National Institute for Research and Development in Informatics
8-10 Averescu Blvd., Bucharest, Romania, badea@ici.ro

Existing clustering methods do not deal well with overlapping clusters, are unstable and do not take into account the robustness of biological systems, or more complex background knowledge such as regulator binding data. Here we describe a nonnegative sparse factorization algorithm dealing with the above problems: cluster overlaps are allowed by design, the nonnegativity constraints implicitly approximate the robustness of biological systems and regulator binding data is used to guide the factorization. Preliminary results show the feasibility of our approach.

1 Introduction and motivation

The advent of microarray technology has allowed a revolutionary transition from the exploration of the expression of a handful of genes to that of entire genomes. However, despite its enormous potential, microarray data has proved difficult to analyze, partly due to the significant amount of noise, but also due to the large number of factors that influence gene expression (many of which are *not* at the mRNA/transcriptone level) and the complexity of their interactions.

One of the most successful microarray data analysis methods has proved to be *clustering* (of genes and/or samples), and a large variety of such methods have been proposed and applied to real-life biological data. This large body of work, impossible to extensively review here, has emphasized important limitations of existing clustering algorithms:

(1) Most clustering methods produce *non-overlapping* clusters. However, since genes are typically involved in several biological processes, "non-overlapping" clustering methods, such as hierarchical clustering (HC) [2], self-organizing maps (SOM) [12], k-means clustering, etc., tend to be unstable, producing different gene clusters for only slightly different input samples (e.g. in the case of HC), or depending on the choice of initial conditions (as in the case of SOM [5], or k-means).

Algorithms allowing for overlapping clusters, such as fuzzy k-means [4] achieved significant improvements w.r.t. "non-overlapping" clustering, but they still have the problems discussed below.

(2) Most algorithms perform clustering along a single dimension comparing e.g. genes w.r.t. *all* the available samples, whereas in reality genes have coordinated expression levels only for certain subsets of conditions. Algorithms dealing with this problem, such as biclustering [13], coupled-two way clustering (CTWC) [3], ISA (iterative signature) [1] have other problems mostly related to the control of overlap between biclusters.

(3) Although genes are subject to both positive and negative influences from other genes, the *robustness* of biological systems requires that an observed change in the expression level of a given gene is the result of *either* a positive *or* a negative influence rather than a complex combination of positive and negative influences that partly cancel out each other (as in the case of Principal Component Analysis).

Nonnegative Matrix Factorization (NMF) [9] deals with this problem by searching for *nonnegative* decompositions of (nonnegative) data. The observed localized nature of the decompositions seems to be a biproduct of the nonnegativity constraints [9].

Recently, Brunet at al [5] applied NMF for clustering samples in a *non-overlapping* mode for three cancer datasets. While oligonucleotide arrays used in that work produce *positive* data, which lend themselves naturally to nonnegative decompositions, clustering genes in an analogous manner would ignore potential negative influences (i.e. genes downregulating other genes).

On the other hand, Kim and Tidor [6] used NMF to cluster genes in the context of a large dataset of yeast perturbation experiments (spotted arrays) [7]. Although NMF has the tendency of producing sparse representations, the factorizations obtained were subjected to thresholding and subsequent reoptimization to obtain sufficiently sparse clusters.

Unfortunately however, microarray data is noisy and it might be useful to be able to take into account any background knowledge that may be available. For example, Lee et al [8] have published binding (location analysis) data for a large number (106) of transcription factors in the yeast S. Cerevisiae.

Elsewhere [in preparation] we have observed that transcription factor (TF) expression levels are not always good predictors for the expression levels of their targets. (The presence of the transcription factor is of course required for the target to be expressed, but very frequently the TF is activated by a different signaling molecule, e.g. a kinase.) Therefore using the binding data directly as background knowledge may not be very helpful in practice.

In fact, it seems that although TFs are not well correlated with their targets, the targets themselves seem to be much better correlated among each other.

In the following, we show how TF binding data can be used as background knowledge in a novel nonnegative factorization algorithm $NNSC_B$ (Nonnegative Sparse Coding with Background Knowledge) designed to address the above-mentioned problems of existing clustering algorithms.

Our algorithm is an improvement of the nonnegative sparse coding (NNSC) algorithm of Hoyer [11]. It produces overlapping clusters that are much more stable than those generated with other algorithms, while also being able to take background knowledge into account.

2 The data sources

Since the most extensive background knowledge is available for the yeast S. Cerevisiae, in this paper we use the Rosetta Compendium [7], the largest publicly available gene expression dataset for yeast perturbations, as well as the binding (location analysis) data of Lee et al. [8].

The Rosetta Compendium contains expression profiles over virtually all yeast genes (6315 ORFs) corresponding to 300 diverse mutations and chemical treatments (276 deletion mutants, 11 tetracycline-regulatable alleles of essential genes and treatments with 13 well-characterized compounds) of S. Cerevisiae grown under a single (normal) condition. The data contains log-expression ratios $\log_{10} r = \log_{10} \frac{g(tf\Delta)}{g(WT)}$, where $g(tf\Delta)$ and $g(WT)$ are the mRNA concentrations of gene g in the $tf\Delta$ mutant and the wild type respectively.

The location analysis data of Lee et al. contains information about binding of 106 transcriptional regulators to upstream regions of target genes.

Since log-ratios can be negative, we cannot directly apply a nonnegative factorization algorithm to the log-ratio dataset. On the other hand, although the ratios r (or maybe r-1) *are nonnegative,* applying nonnegative factorization on them would only uncover the positive influences, while in practice the low level of certain genes is due to them being *downregulated* by other genes.

To address problem (3) mentioned in the Introduction, we separate, as in [6], the up-regulated from the down-regulated part of each gene, i.e. obtain two entries g^+ and g^- for each gene g from the original gene expression matrix:

$$r(g^+) = \begin{cases} r(g) - 1 & r(g) \geq 1 \\ 0 & otherwise \end{cases} \qquad r(g^-) = \begin{cases} \frac{1}{r(g)} - 1 & r(g) < 1 \\ 0 & otherwise \end{cases}$$

Note that in this representation, a significantly downregulated gene will require a non-negligible contribution in the factorization. (We use the ratios rather than log-ratios as in [6], since linear combinations of log-ratios amount to products of powers of ratios rather than additive contributions.)

3 Nonnegative Sparse Coding

Hoyer's NNSC algorithm [11] factorizes a $n_s \times n_g$ matrix $X \approx A \cdot S$ as a product of an $n_s \times n_c$ matrix A and a $n_c \times n_g$ matrix S by optimizing (minimizing) the following objective function: [1]

[1] To achieve information compression, the number of internal dimensions n_c must verify the constraint $n_c(n_s+n_g) < n_s n_g$.

$$C(A,S) = \frac{1}{2}\left\|X - AS\right\|_F^2 + \lambda \sum_{c,g} S_{cg} \qquad (1)$$

with respect to the *nonnegativity constraints* $A_{sc} \geq 0,\ S_{cg} \geq 0$ (2)

The objective function combines a *fitness* term involving the Frobenius norm of the error and a *size* term penalizing the non-zero entries of S. (The Frobenius norm is given by $\left\|E\right\|_F^2 = \sum_{s,g} E_{sg}^2 = Tr(E^T E).$)

The *Nonnegative Matrix Factorization* (NMF) of Lee and Seung [10] is recovered by setting the size parameter λ to zero, while a non-zero λ would lead to sparser factorizations.

The objective function (1) above has an important problem, due to the invariance of the fitness term under diagonal scalings. More precisely we have the following result.

Proposition. The fitness term $\frac{1}{2}\left\|X - AS\right\|_F^2$ (i.e. the NMF objective function) is invariant under the following transformations:

$$A \leftarrow A \cdot D^{-1}, \quad S \leftarrow D \cdot S, \qquad (3)$$

where $D = \mathrm{diag}(d_1,\ldots,d_{nc})$ is a positive diagonal matrix ($d_c > 0$).

Note that such positive diagonal matrices are the most general positive matrices whose inverses are also positive (thereby preserving the nonnegativity of A and S under the above transformation).

The scaling invariance of the fitness term in (1) makes the size term ineffective, since the latter can be forced as small as needed simply by using a diagonal scaling D with small enough entries. Additional constraints are therefore needed to render the size term operational. Since a diagonal matrix D operates on the rows of S and on the columns of A, we could impose unit norms either for the rows of S, or for the columns of A.

Unfortunately, the objective function (1) used in [11] has an important flaw: it produces decompositions that depend on the scale of the original matrix X (i.e. the decompositions of X and αX are essentially different), regardless of the normalization scheme employed. For example, if we constrain the rows of S to unit norm, then we cannot have decompositions of the form $X \approx A \cdot S$ and $\alpha X \approx \alpha A \cdot S$, since at least one of these is in general non-optimal due to the dimensional inhomogeneity of the objective function w.r.t. A and X

$$C_{\alpha X}(\alpha A, S) = \alpha^2 \frac{1}{2}\left\|X - AS\right\|_F^2 + \lambda \sum_{c,g} S_{cg}\,.$$

On the other hand, if we constrain the columns of A to unit norm, the decompositions $X \approx A \cdot S$ and $\alpha X \approx A \cdot \alpha S$ cannot be both optimal, again due to the dimensional inhomogeneity of C, now w.r.t. S and X:

$$C_{\alpha X}(A, \alpha S) = \alpha^2 \frac{1}{2}\|X - AS\|_F^2 + \alpha\lambda \sum_{c,g} S_{cg}$$

Therefore, as long as the size term depends only on S, we are forced to constrain the columns of A to unit norm, while employing an objective function that is *dimensionally homogeneous* in S and X. One such dimensionally homogeneous objective function is:

$$C(A, S) = \frac{1}{2}\|X - AS\|_F^2 + \lambda\|S\|_F^2 \qquad (1')$$

which will be minimized w.r.t. the nonnegativity constraints (2) and the constraints on the norm of the columns of A:

$$\|A_c\| = 1 \qquad \text{(i.e. } \sum_s A_{sc}^2 = 1\text{)} \qquad (4)$$

It can be easily verified that this produces *scale independent decompositions*, i.e. if $X \approx A \cdot S$ is an optimal decomposition of X, then $\alpha X \approx A \cdot \alpha S$ is an optimal decomposition of αX.

The constrained optimization problem could be solved with a gradient-based method. However, in the case of NMF, faster so-called "multiplicative update rules" exist [10,11], which we have generalized to the NNSC problem as follows. (These methods only produce local minima, but the solutions tend to be quite 'stable' – see also Section 5 below.)

Modified NNSC algorithm

Start with random initial matrices A and S

loop

$$S_{cg} \leftarrow S_{cg} \frac{(A^T \cdot X)_{cg}}{(A^T \cdot A \cdot S + \lambda S)_{cg}}$$

$$A \leftarrow A + \mu(X - A \cdot S) \cdot S^T$$

normalize the columns of A to unit norm: $A \leftarrow A \cdot D^{-1}$, $D = diag\left(\sqrt{\sum_s A_{sc}^2}\right)$

until convergence.

In the following, we assume that the gene expression data is given in an $n_s \times n_g$ matrix X, where n_s and n_g are the numbers of samples and genes respectively, so that X_{sg} represents the expression level of gene g in sample s.

A sparse factorization $X \approx A \cdot S$ will be interpreted as a generalization of *clustering the genes* (i.e. the columns X_g of X) into overlapping clusters c corresponding to the rows S_{cg} of S. More precisely, a non-zero value of S_{cg} (or at least an S_{cg} larger than a given threshold) will be interpreted as the gene g belonging to cluster c. Note that clusters can be overlapping, since the columns of S may have several significant entries.

Although overlaps are allowed, NNSC will not produce highly overlapping clusters, due to the sparseness constraints. This is unlike many other clustering

algorithms that allow clusters to overlap, which have to resort to several parameters to keep excessive cluster overlap under control.)

Also note that we factorize X rather than X^T since the sparsity constraint should affect the clusters of genes (i.e. S) rather than the clusters of samples A. (This is unlike NMF, for which the factorizations of X and of X^T are completely symmetrical.)

4 Nonnegative Sparse Coding using Background Knowledge

Transcription factor binding data can be represented by a $n_f \times n_g$ Boolean matrix B, such that $B_{fg}=1$ iff the transcription factor f binds to the upstream region of gene g.

As already mentioned in the Introduction, although transcription factor expression levels are not always good predictors of the expression levels of their targets, the targets are frequently much better correlated among themselves. This suggests using the co-occurrence matrix $K=B^TB$ rather than B as background knowledge. K is an $n_g \times n_g$ square matrix, in which $K_{g'g''} \neq 0$ iff genes g' and g'' are both targets of some common transcription factor f.

Our idea of exploiting background knowledge during clustering is quite simple. Normally, non-zero entries in the "gene cluster" matrix S are penalized for size. However, if certain entries conform to the background knowledge, they will be exempted from size penalization. We thus need to modify the size term in $C(A,S)$ to take into account B.

Implementing this simple idea involves however certain subtleties. Assume that some gene cluster c (i.e. set of genes g for which $S_{cg} \neq 0$, or at least $S_{cg} > T$ for some given threshold T) contains many genes that are targets of several TFs (e.g. Figure 1b below). Although this cluster is preferable from the point of view of the background knowledge to the one from Figure 1a, it is worse than the one from Figure 1c, in which all the genes are the targets of a single TF.

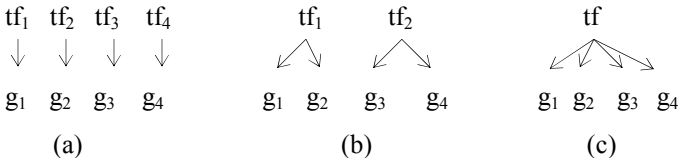

Figure 1. Conformance of clusters to the binding data: (c) is preferable to (b), which is preferable to (a).

Thus, the size term cannot be simply a sum of overlaps of the clusters (i.e. rows of S) with groups of TF-targets (rows of B), since

$$\sum_c \sum_f \sum_g S_{cg} B_{fg} = \sum_g \left(\sum_c S_{cg}\right)\left(\sum_f B_{fg}\right)$$

does not depend on the way the genes are distributed in groups of TF-targets for different TFs.

Clusters like the one in Figure 1c can be highly evaluated if cross-terms between genes controlled by the same TF are added. We thus encourage genes controlled by the same TF in the binding data, while penalizing the size of S using an objective function of the form:

$$C(A,S) = \frac{1}{2}\|X - AS\|_F^2 + \frac{\lambda}{2}\left(\|S\|_F^2 - Tr(SKS^T)\right) \quad (5)$$

with $K = \gamma \cdot \left(\sigma(B^T B) - diag\left(\sigma(\sum_f B_{fg})\right)\right)$, where γ is a factor such that $\frac{1}{n_g}\sum_{g',g''} K_{g'g''} = 1$ and $\sigma(\cdot)$ is the Heaviside step function (applied element-wise). Of course, optimization of (5) is attempted in the context of the constraints (2) and (4). The algorithm below solves the above optimization problem using combined multiplicative and additive update rules. (The final normalization of the rows of S renders the resulting clusters comparable.)

NNSC$_B$ algorithm

start with random initial matrices A and S

loop

$$S_{cg} \leftarrow S_{cg}\frac{(A^T \cdot X + \lambda S \cdot K)_{cg}}{(A^T \cdot A \cdot S + \lambda S)_{cg}}$$

$$A \leftarrow A + \mu(X - A \cdot S) \cdot S^T$$

normalize the columns of A to unit norm: $A \leftarrow A \cdot D^{-1}$, $D = diag\left(\sqrt{\sum_s A_{sc}^2}\right)$

until convergence

$S \leftarrow D^{-1} \cdot S$ and $A \leftarrow A \cdot D$, where $D = diag\left(\sqrt{\sum_g S_{cg}^2}\right)$

To test our approach, we have applied the NNSC$_B$ algorithm on a synthetic dataset with several highly overlapping clusters. NNSC$_B$ has been able to consistently recover the clusters, or close approximations thereof even in the presence of noise. (See http://www.ai.ici.ro/psb05/synthetic.pdf for more details.)

5 Clustering the Rosetta dataset w.r.t. the binding data of Lee et al.

Although the main goal of this paper is the presentation of a new clustering algorithm able to deal with background knowledge rather than obtaining new biological insights, we also briefly discuss our initial attempts at applying our algorithm to yeast microarray data.

The binding data of Lee et al. contains the targets of 106 transcription factors, roughly about half the total number of yeast transcription factors. In order not to introduce a bias towards the targets of these TFs due to the background knowledge,

we have selected from the Rosetta dataset only these targets. (We also eliminated the genes that had unreliable measurements in the Rosetta dataset – dealing with missing values in our context is a matter of future work.) This left us with a set of 99 TFs and 2099 genes. The matrix X to be factorized was constructed by duplicating genes as described in Section 2 (X has therefore 4198 columns). Duplicating genes g into their positive (g^+) and negative parts (g^-) may raise potential problems with possible conflicts between nonzero S_{cg+} and S_{cg-} entries, as a gene cannot be *both* up- and down regulated in a given cluster. The fact that our decompositions never have both S_{cg+} and S_{cg-} nonzero (significant) shows that the approach is biologically sensible.

An important parameter of the $NNSC_B$ factorization is its *internal dimensionality* (the number of clusters n_c). A useful estimate of the internal dimensionality of a dataset can be obtained from its singular value decomposition (SVD).

A more refined analysis [6] determines the number of dimensions around which the root mean square error (RMSE) *change* of the real data and that of a randomized dataset become equal. Kim and Tidor's analysis estimated the internal dimensionality of the Rosetta dataset to be around 50. We performed this analysis for our restricted dataset and obtained a similar dimensionality around 50 (see Figure 2).

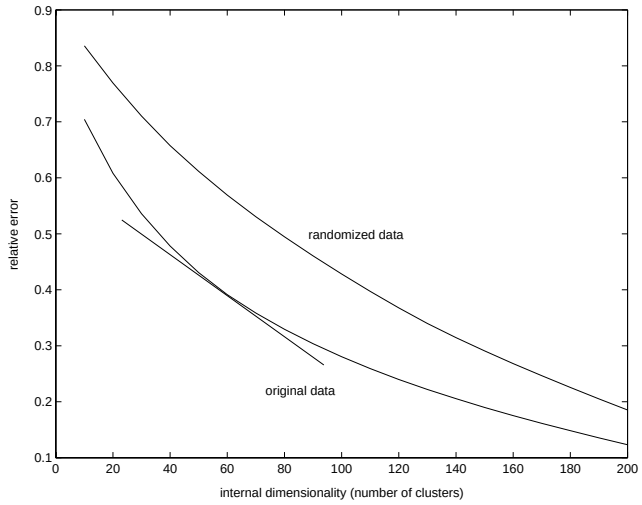

Figure 2. Determining the internal dimensionality of the Rosetta dataset using the method of Kim and Tidor

We also considered another approach: for a given set of clusters we determined the fraction of *intra-cluster* correlated pairs of genes (i.e. the number of correlated pairs of genes divided by the total number of correlated pairs of genes). This fraction was estimated for various correlation thresholds r_0. (More precisely, a gene pair (g_1,g_2) is called correlated w.r.t. threshold r_0 iff $| r(g_1,g_2) | \geq r_0$.) Figure 3 depicts

the r_0-dependence of the fraction of intra-cluster correlated pairs of genes for various dimensionalities. Notice that n_c=50 is a reasonable choice as the above mentioned fraction approaches 90% for a large range of r_0. (This means in other words that most of the correlated gene pairs are *within* clusters.)

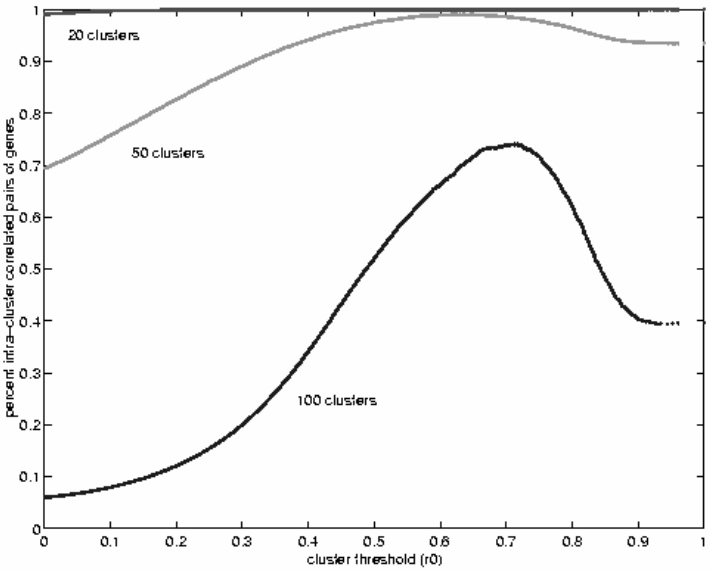

Figure 3. The correlation-threshold dependence of the fraction of intra-cluster correlated gene pairs.

Next we studied the *stability* of the algorithm w.r.t. the initial starting point. (The algorithm was run disregarding the background knowledge, i.e. taking λ=0, in order to avoid any possible influences of the background knowledge on the stability of some solution fragment.) The Table below lists the relative errors ($\|X-AS\|_F/\|X\|_F$) obtained in 7 runs for 7 different initializations.

run no.	1	2	3	4	5	6	7
relative error	0.1544	0.1677	0.1636	0.1718	0.1592	0.1606	0.1782

To asses the stability of the decomposition in the case of k runs, we determined the best matches among the k sets of clusters. The following Table shows the numbers of matching clusters for a progressively larger number of runs. Note the (relative) stabilization of the number of matching clusters w.r.t. the numbers of runs:

Numbers of runs	1	2	3	4	5	6	7
Matching clusters (avg.)	50	48.4	42.2	40.1	37.2	36	35

Next, we studied the influence of the background knowledge on the solution. In order to separate the variability of the results due to the initialization from that due

to the background knowledge, we performed all the following tests using the same initialization. We ran $NNSC_B$ with several values of the parameter λ, ranging from 0 (background knowledge is not taken into account) to 0.75 (background knowledge has comparable weight to the data fit term) and tested the overlap of the resulting clusters with the background knowledge. Briefly, we observed a clear increase of this overlap with larger λ. More precisely (more details can be found in the Table below):

- the average fraction of cluster genes controlled by TFs increased from about 47% for λ=0 to 66% for λ=0.75. (Only the TFs controlling at least two genes from the cluster were counted.)
- the average number of TFs per cluster controlling at least two genes increased from around 8 for λ=0 to 14 for λ=0.75.

λ	0	0.1	0.2	0.5	0.75
avg. number of genes in a cluster	27.8	32	36.4	43.6	78.7
avg. fraction of cluster genes controlled by TFs	0.47	0.46	0.51	0.60	0.66
avg. number of TFs per cluster	7.9	8.6	10.4	13.4	17.9
avg. cluster overlap	1	1.3	2.1	3.6	10
avg. cluster overlap (when only overlapping pairs are averaged)	2.7	3.3	4.3	6.2	14.7
Relative error	0.16	0.18	0.23	0.33	0.41

Of course, conformance to the data and/or to the background knowledge does not imply the biological relevance of the results. To estimate the latter, we searched for significant Gene Ontology (GO) annotations [14] of the genes in the clusters. More precisely, we employed the hypergeometric distribution to compute p-values representing likelihoods that specific GO annotations and a given cluster share a given number of genes by chance and retained only the annotations with a p-value less than 10^{-3}. (This p-value threshold was chosen so that not more than 1 or 2 annotations are false positives, given that the genes in our dataset share 2422 GO annotations – some of which may of course be related by sub- or super-class relationships. We did not use a lower threshold in order to avoid many missing annotations.)

To demonstrate the biological relevance of the factorizations using background knowledge, we performed alternative factorizations for randomized background knowledge (more precisely, by randomly permuting the lines of B independently of each other). This lead to a drop in the average number of significant GO annotations per cluster from 8.16 to 4.88 (for λ=0.75).

We also looked at a few clusters in more detail. For example, cluster 47 had 18 genes, involved in the STE12 control of pheromone response, among which, for instance, AGA1, FIG1, FUS1 are involved in cell fusion, while GPA1, FUS3 and PRR2 are involved in the pheromone signal transduction pathway, KAR4 is a regulatory protein required for pheromone induction of karyogamy genes and SST2 is involved in desensitization to alpha-factor pheromone. The mating a-factor genes MFA1 and MFA2 are also in the cluster. The entire cluster is presented in the

Annex, together with the associated significant GO annotations and the cluster coefficients from the S matrix. (The threshold used for extracting clusters from the S matrix was $1/\sqrt{n_g} = 0.0154$.)

5 Conclusions

Despite their wide-spread use in microarray data analysis, existing clustering algorithms have serious problems, the most important one being related to the fact that biological processes are overlapping rather than isolated. The impact of microarray technology is also limited by the noisy nature of measurements, which can only be compensated by additional background knowledge. Here we have shown how these important problems faced by microarray data analysis can be dealt with in the context of a sparse factorization algorithm capable of dealing with regulator binding data as background knowledge. A key ingredient of this algorithm is the nonnegativity constraint. Such an approach, for example using NMF, has been mostly advocated in connection with oligonucleotide array data, which are (at least theoretically) nonnegative. However, this viewpoint is only partially correct, since downregulated genes would not be explained in such a framework. Actually, we argue that nonnegative factorizations are appropriate due to the robustness of biological systems, in which an observed change in a gene's expression level is the result of *either* a positive *or* a negative influence rather than a complex combination of the two. Although our preliminary results are encouraging, a more detailed biological analysis should be the focus of subsequent work. (The clusters obtained by our algorithm for various parameter settings can be found online at http://www.ai.ici.ro/psb05/.)

References

1. Bergmann S., J. Ihmels, N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review E 67, 031902 (2003).
2. Eisen M.B., P.T. Spellman, P.O. Brown, D. Botstein. Cluster analysis and display of genome-wide expression patterns, PNAS Vol.95, 14863-8, Dec.1998.
3. Getz G., Levine E., Domany E. Coupled two-way clustering analysis of gene microarray data, PNAS Vol. 97, 12079-84, 2000.
4. Gasch A.P, Eisen M.B. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol. 2002, Oct 10;3(11).
5. Brunet J.P., Tamayo P., Golub T.R., Mesirov J.P. Metagenes and molecular pattern discovery using matrix factorization. PNAS 101(12):4164-9, 2004, Mar 23.
6. Kim P.M., Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 2003 Jul;13(7):1706-18.
7. Hughes T.R. et al. Functional discovery via a compendium of expression profiles. Cell. 2000 Jul 7;102(1):109-26.

8. Lee T.I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science. 2002 Oct 25; 298(5594):799-804.
9. Lee D.D., H.S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, vol. 401, no. 6755, pp. 788-791, 1999.
10. Lee D.D., H.S. Seung. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing 13 (Proc. NIPS*2000), MIT Press, 2001.
11. P. O. Hoyer. Non-negative sparse coding. Neural Networks for Signal Processing XII, 557-565 Martigny, Switzerland, 2002.
12. Tamayo P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, PNAS 96(6):2907-12 (1999).
13. Cheng Y, Church GM. Biclustering of expression data. Proc. ISMB 2000; 8:93-103.
14. Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. 25:25-29, 2000.

Annex. The 18 genes of cluster 47 and the TFs controlling them

Significant GO annotations (with p-values): conjugation(0),conjugation with cellular fusion(0),sexual reproduction(0),reproduction(2.30e-11),response to pheromone(4.96e-11),response to chemical substance(2.24e-9),response to abiotic stimulus(2.89e-9),development(1.07e-8),response to pheromone during conjugation with cellular fusion 1.13e-8),response to external stimulus(1.34e-8),cell communication(2.51e-8),signal transduction during conjugation with cellular fusion(1.57e-6),response to stimulus(2.11e-6),signal transduction(3.85e-6),G-protein coupled receptor protein signaling pathway(4.27e-6),cell surface receptor linked signal transduction(9.42e-6),shmoo tip(9.42e-6),signal transducer activity(4.42e-5),pheromone activity(1.97e-4),receptor binding(1.97e-4),cellular process(3.05e-4),receptor signaling protein serine/threonine kinase activity(3.92e-4),site of polarized growth(9.32e-4),site of polarized growth (sensu Fungi)(9.32e-4),site of polarized growth (sensu Saccharomyces)(9.32e-4),receptor signaling protein activity(9.70e-4)

STE12 targets: STE12, SST2, TEC1, FUS1, KAR4, GPA1, MFA2
MCM1 targets: MFA1, STE6, MFA2, AGA1
PHD1 targets: PRR2, GPA1

+: 0.000000	-: 0.970648	PRR2	strong similarity to putative protein kinase NPR1
+: 0.000000	-: 0.146131	YDL133W	
+: 0.000000	-: 0.132290	FUS1	cell fusion protein
+: 0.064781	-: 0.000229	PRY3	
+: 0.000000	-: 0.056725	FIG1	required for efficient mating
+: 0.000000	-: 0.040888	AGA1	a-agglutinin anchor subunit
+: 0.029486	-: 0.000000	HSP12	heat shock protein
+: 0.000000	-: 0.027441	MFA2	mating pheromone a-factor 2
+: 0.000000	-: 0.023362	GPA1	GTP-binding protein alpha subunit of the pheromone pathway
+: 0.000000	-: 0.022420	questionable ORF	
+: 0.000062	-: 0.022318	STE6	ATP-binding cassette transporter protein
+: 0.000000	-: 0.020931	STE12	transcriptional activator
+: 0.000000	-: 0.020907	FUS3	mitogen-activated protein kinase (MAP kinase)
+: 0.000000	-: 0.020485	MFA1	mating pheromone a-factor 1
+: 0.000000	-: 0.020241	SST2	involved in desensitization to alpha-factor pheromone
+: 0.000000	-: 0.020059	YLR343W	
+: 0.000000	-: 0.019342	TEC1	Ty transcription activator
+: 0.000103	-: 0.018020	KAR4	regulatory protein required for pheromone induction of karyogamy genes

Parameters: λ=0.1, $\mu=5\cdot10^{-6}$, *S*-threshold=0.0154, 300 iterations.